Received: 18 December 2021 | Accepted: 22 March 2022

DOI: 10.1111/2041-210X.13858

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Estimating species misclassification with occupancy dynamics and encounter rates: A semi-supervised, individual-level approach

Anna I. Spiers[1,2] | J. Andrew Royle[3] | Christa L. Torrens[4] | Maxwell B. Joseph[1]

[1]Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA

[2]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA

[3]U.S. Geological Survey, Eastern Ecological Science Center, Laurel, MD, USA

[4]Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, USA

Correspondence
Anna I. Spiers
Email: anna.spiers@colorado.edu

Funding information
University of Colorado CIRES Earth Lab

Handling Editor: Miguel Acevedo

Abstract

1. Large-scale, long-term biodiversity monitoring is essential to conservation, land management and identifying threats to biodiversity. However, multispecies surveys are prone to various types of observation error, including false-positive/false-negative detection and misclassification, where a species is thought to have been encountered but not correctly identified. Previous methods assume an imperfect classifier produces species-level classifications, but in practice, particularly with human observers, we may end up with extraspecific classifications including 'unknown', morphospecies designations and taxonomic identifications coarser than species. Disregarding these types of species misclassification in biodiversity monitoring datasets can bias estimates of ecologically important quantities such as demographic rates, occurrence and species richness.
2. Here we present a joint classification-occupancy model that accounts for species non-detection and misclassification. Our framework accommodates extinction and colonization dynamics, allows for additional uncertain 'morphospecies' designations and makes use of individual specimens with known species identities in a semi-supervised setting. We compare the performance of our model to a classification-only model that discards information about occupancy and encounter rate. We illustrate our model with an empirical case study of the carabid beetle (Carabidae) community at the National Ecological Observatory Network Niwot Ridge Mountain Research Station, near Boulder, CO, USA. We also use simulations to evaluate model performance through validation metrics where varying fractions of the data are confirmed.
3. The model supported imperfect classifier accuracy and favoured certain true species classifications strongly for some morphospecies. The model outperformed (e.g. precision) the reduced model that discarded occupancy information, and these differences were most pronounced for abundant species.
4. Spatial and temporal dynamics from modelled occupancy and encounter rates may inform species misclassification probability, but this idea has not yet been

tested. Our statistical framework explores this opportunity, and can be applied to datasets with imperfect species detection and classification, limited verification data and non-species classifications.

KEYWORDS
carabid, imperfect classifier, morphospecies, NEON, observation error, occupancy models, semi-supervised, species misclassification

1 | INTRODUCTION

Large-scale, long-term biodiversity monitoring is essential to conservation and land management and identifying threats to biodiversity. Such comprehensive datasets increasingly include multispecies surveys that capture information-rich co-occurrence data, enabling community-level analyses (Iknayan et al., 2014; Ovaskainen et al., 2017). However, multispecies surveys are prone to various types of errors, including false absences where a species is present but not detected (Dorazio & Royle, 2005), and misidentification, where a species is encountered but its species identity is not correctly recorded (Miller et al., 2011).

Certain classes of occupancy models account for observation error in biodiversity surveys that seek to understand species distributions, track population changes and describe mechanisms underlying population and community dynamics (MacKenzie et al., 2002). Latent presence/absence states are modelled explicitly, with an observation model that accounts for the details of the detection process, including the potential for false negatives (non-detections at occupied sites) and false positives (detections at unoccupied sites; Chambert et al., 2015; Miller et al., 2012; Royle & Link, 2006; Wright et al., 2020). Disregarding false positives in biodiversity monitoring data can bias estimates of ecologically important quantities such as demographic rates, occurrence and species richness (Chambert et al., 2015, 2018; McClintock et al., 2010).

Multi-species surveys are also subject to errors in species identifications by imperfect classifiers. Imperfect classifiers include citizen scientists (e.g. North American Breeding Bird Survey; Sauer et al., 2017), technicians trained in local taxonomy (e.g. invertebrate trapping by NEON; Hoekman et al., 2017), automated methods (e.g. bat acoustic recording software; Wright et al., 2020) or convolutional neural networks used with camera trap data; Tabak et al., 2019). Previous methods assume an imperfect classifier produces species-level classifications, but in practice, particularly with human observers, we may end up with extraspecific classifications including 'unknown', morphospecies designations (i.e. individuals that cannot be taxonomically identified so are grouped by morphology) and taxonomic identifications coarser than species.

If species observations are prone to misclassification, then samples with verified species identities might be used to estimate misclassification probabilities. We refer to this situation as 'semi-supervised': true species identities are known for some but not all individuals. However, leveraging these partially observed, individual-level validation data for the rest of the dataset present a methodological challenge. Previous multi-species occupancy models that accommodate misclassification have used site-level validation data where the occupancy state of a species is known only at a site but not at an individual sample level (Chambert et al., 2018) or multinomial models with site-level covariates that aggregate individual samples (Wright et al., 2020). Using an individual, sample-level approach can help in resolving non-species (e.g. morphospecies) identities to the true species identity.

Misclassified species identities can be dealt with using one of two contrasting approaches. A simple two-step approach in which (a) a classifier is used to assign species identities to each individual (creating one complete synthetic dataset from classifier output, for which species identities are treated as known or are verified using an unambiguous classification method), then (b) the synthetic dataset is analysed using a downstream model (e.g. an occupancy model). This two-stage approach does not use any information about occupancy or encounter rates in the first stage. An alternative approach is to simultaneously model the classification process and the ecological process in a single joint model. A joint model directly uses classifier output as data, relating the observation process to underlying, imperfectly observed, ecological states in one step. Such an approach can simultaneously account for uncertainty in species identities, and use information about occupancy and encounter rates to inform species identity estimates (Wright et al., 2020). However, there remains the practical question of how much value is added by a joint model vs. a two-stage approach. A priori, we expect that a joint model should produce better true species identities simultaneously by directly modelling the link between ecological states and the observation process, but this has not yet been tested.

Here we present an individual-level, semi-supervised, dynamic occupancy model that accounts for species non-detection and misclassification. Our Bayesian approach extends the classification-occupancy model of Wright et al. (2020) to (a) accommodate extinction and colonization dynamics, (b) allow for additional uncertain morphospecies designations in the imperfect species classifications and (c) make use of labelled samples with known species identities in a semi-supervised setting. Furthermore, we compare the classification performance (i.e. accuracy and precision of posterior draws) of a joint classification-occupancy model to a reduced classification-only model that discards information about occupancy and encounter rate on a withheld test set. We demonstrate our model using simulations and with an empirical case study of the

carabid beetle (Carabidae) community at the National Ecological Observatory Network (NEON) Niwot Ridge Mountain Research Station (NIWO), west of Boulder, CO, USA.

2 | MATERIALS AND METHODS

2.1 | Modelling occupancy dynamics and misclassification

Consider data collected at sites $i = 1, \ldots, N$, according to a robust design (Hoekman et al., 2017) where each site is visited $j = 1, \ldots, J$ times within primary periods $t = 1, \ldots, T$, where the occupancy states are assumed to be static within primary periods. We are interested in occupancy states (true presence or absence) and encounter rates (observed frequency) for species $k = 1, \ldots, K$.

2.1.1 | State model

We assume that the objective of classification is to use the resulting data in an ecological model describing species occurrence over space and time. Let $z_{i,k,t}$ be the binary occurrence state for species $k = 1, \ldots, K$ at site i and during time t. Sites are either occupied ($z_{i,k,t} = 1$) or not ($z_{i,k,t} = 0$). We assume that the occupancy states arise as Bernoulli random variables:

$$z_{i,k,t} \sim \text{Bernoulli}\left(\psi_{i,k,t}\right).$$

The probability of occurrence in the initial primary period is $\psi_{i,k,1}$. Subsequent occupancy dynamics depend on the probability of colonization $\gamma_{i,k,t}$ and persistence $\phi_{i,k,t}$, such that for $t > 1$:

$$\psi_{i,k,t} = z_{i,k,t-1}\phi_{i,k,t-1} + \left(1 - z_{i,k,t-1}\right)\gamma_{i,k,t-1}.$$

2.1.2 | Observation model: Encounters

On any particular sampling occasion j at site i in period t, we encounter $L_{i,j,k,t}$ individuals with encounter rate $\lambda_{i,j,k,t}$. We assume that the number of encounters is a Poisson random variable: $L_{i,j,k,t} \sim \text{Poisson}\left(z_{i,k,t}\lambda_{i,j,k,t}\right)$. So $L_{i,j,k,t} = 0$ indicates non-detection, due to either the species not occupying that site (i.e. $z_{i,k,t} = 0$) or the species may occupy the site but was not encountered (i.e. $\lambda_{i,j,k,t} = 0$). In a setting with misclassification, the number of encountered individuals $L_{i,j,k,t}$ is not observed directly because of uncertainty in the true species identities of encountered individuals. We do however observe the total number of individuals across all species encountered on any particular occasion: $L_{i,j,.,t} = \sum_{k=1}^{K} L_{i,j,k,t}$. The properties of sums of Poisson random variables allow us to model these observed totals as:

$$L_{i,j,.,t} \sim \text{Poisson}\left(\sum_{k=1}^{K} z_{i,k,t}\lambda_{i,j,k,t}\right).$$

2.1.3 | Observation model: Classification

In addition to observing the total number of encountered individuals on an occasion, $L_{i,j,.,t}$, we assume that we also obtain imperfect species classifications for each encountered individual. In cases where individuals have been encountered ($L_{i,j,.,t} > 0$), we obtain imperfect classifications of individuals $l = 1, \ldots, L_{i,j,.,t}$ and model these as arising from a categorical distribution with a species-specific probability vector:

$$y_{i,j,l,t} \sim \text{Categorical}\left(\theta_{k[i,j,l,t]}\right),$$

where $y_{i,j,l,t}$ is the imperfect classification, and $\theta_{k[i,j,l,t]}$ is a probability vector associated with the true species identity of individual l, which we denote $k[i,j,l,t]$. Element k' in the vector $\theta_{k[i,j,l,t]}$ represents the probability that an individual is classified into category k', conditional on the true species identity $k[i,j,l,t]$, such that $\theta_{k[i,j,l,t],k'} = \Pr\left(y_{i,j,l,t} = k' \mid k[i,j,l,t]\right)$. If species are always misclassified as alternative species (i.e. not morphospecies nor other groups), then θ_k will be a vector of length K (Wright et al., 2020), whereas if there are extraspecific classes (e.g. morphospecies), θ_k may have more than K elements.

True species identities are modelled as:

$$k[i,j,l,t] \sim \text{Categorical}\left(\frac{z_{i,k,t}\lambda_{i,j,k,t}}{\sum_k z_{i,k,t}\lambda_{i,j,k,t}}\right).$$

The occupancy state model is essentially an informative prior on classification probabilities so that where there is a high probability of occurrence of a species, then samples are more likely to be that species compared to when occupancy probability is low.

If ground-truthed species identity data are available for some individuals, then $k[i,j,l,t]$ is partly observed and this model can be used in a semi-supervised setting. In an unsupervised setting, this individual-level formulation is a disaggregated version of the single-season multinomial model of (Wright et al., 2020; Appendix S1). An aggregated version would pool the counts so no identifying information is attributed to any single sample. The disaggregated, or sample-level, approach facilitates the treatment of non-species (e.g. morphospecies) identities and allows for covariates to be included in the model at the individual/observation level, which could improve the estimation of classification probabilities and possibly ecological parameters. For example, sample confidence (i.e. the observer's confidence in species classification for an individual) is a sample-level covariate (like the observation-level covariate, sample quality, in Augustine et al. (2020)) that might be correlated with how samples might be prioritized for verification. This is the case for how samples are prioritized for verification in the NEON dataset, but sample confidence is not available.

2.1.4 | Incorporating morphospecies designations

In some settings, the imperfect classifier might assign more classes than there are unique species so that the vector θ_k has more than K

elements. For example, if an imperfect classifier is unable to identify a set of species, they may classify those individuals as 'unknown' or as a unique morphospecies associated with a given sampling occasion. Thus, it is possible for individuals to be classified into $\tilde{K} \geq K$ classes, where $\tilde{K}$ is the sum of the number of species and the total number of morphospecies designations. In such cases, the matrix $\Theta = (\theta'_1, \ldots, \theta'_K)$ can be rectangular, with the first K columns corresponding to the classification probabilities for species 1, … , K, and the remaining columns corresponding to classification probabilities for non-species (e.g. morphospecies) classes:

$$\Theta = \overbrace{\begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,K} & \cdots & \theta_{1,K'} \\ \vdots & \ddots & & \ddots & \\ \theta_{K,1} & & \theta_{K,K} & & \theta_{K,K'} \end{bmatrix}}^{\substack{\text{Species} \\ \text{classifications}}\quad\quad\substack{\text{Morphospecies} \\ \text{classifications}}} \Bigg\} \begin{matrix} \text{True} \\ \text{species} \\ \text{identities} \end{matrix}$$

2.2 | Case study

2.2.1 | Application to NEON carabid data

We fit our model to the carabid pitfall trap sampling data collected by NEON at NIWO during 2015–2019 (National Ecological Observatory Network, 2021). Carabids are a ubiquitous and speciose family of ground-dwelling invertebrates that are commonly collected by passive sampling methods, like pitfall traps, as described in Hoekman et al. (2017). A well-studied sentinel group, carabids make an excellent study system for assessing community occupancy rates and classification accuracy. Collecting and identifying carabids is resource-intensive, but NEON lowers this barrier to entry by providing a public carabid dataset with three levels of classification (parataxonomist, expert taxonomist, then DNA barcoding). Although NEON processes carabid samples at the domain level (sampling locations within the same ecoregion; Hoekman et al., 2017), we focus our analysis on one NEON sampling location, NIWO, to assess occupancy across co-occurring species. We use the 2015–2019 dataset since carabid sampling started in 2015 at NIWO and expert classification data were not yet fully available for 2020 at the time of analysis due to data latency (National Ecological Observatory Network, 2021). NIWO is a site in the southern Rocky Mountains, spanning subalpine conifer forest and alpine tundra.

We outline the relevant data collection protocol here, but Hoekman et al. (2017) offer more details regarding NEON's carabid pitfall trap data product. The sampling design at every NEON sampling location consists of 10 permanent sites with four pitfall traps per site. Traps are sampled and reset biweekly during the growing season, with a range of 5–7 collections per year at NIWO. In 2018, one site was permanently relocated to ensure sampling was allocated proportionally to the NLCD cover types represented (NEON help desk, personal communication). Variables in our model are defined at the site level.

All carabid samples are classified by a parataxonomist, and a subset are sent to an expert taxonomist for verification (Figure 1; Hoekman et al., 2017). Species classification by parataxonomists is considered imperfect. Identification by an expert taxonomist is treated as confirmation data but is limited due to budget constraints. We confirmed the accuracy of the expert taxonomist classifications in finding that all individuals sent for DNA barcoding by NEON match the expert taxonomist's identification for the samples we used. In

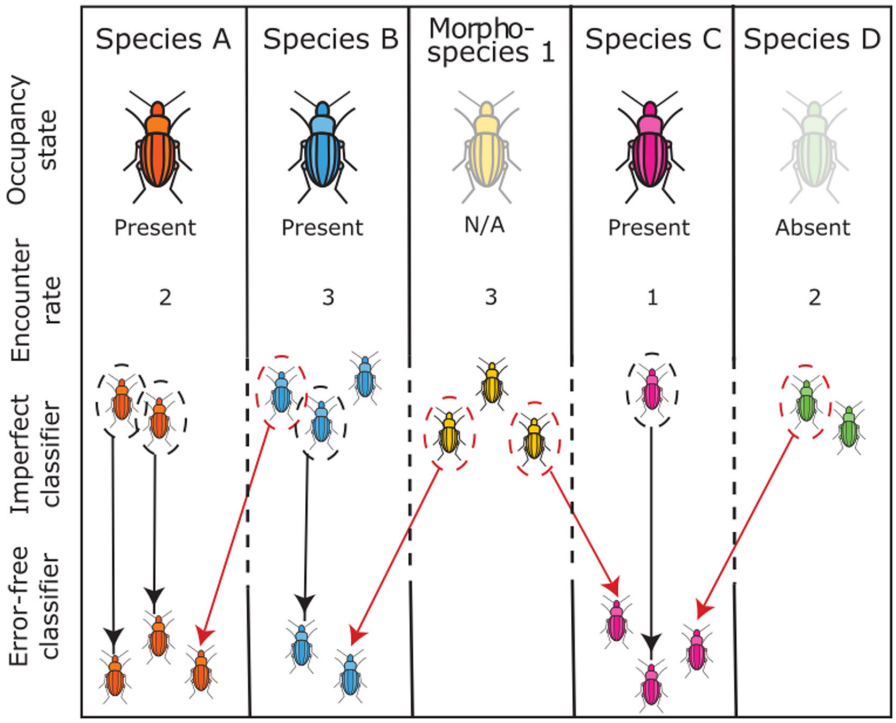

FIGURE 1 Classification scenarios in NEON carabid data. Each column is an imperfect classifier label. Each species is either present or absent, and morphospecies do not have an occupancy state. In some cases, the imperfect classifier (parataxonomist) matches the error-free classifier (expert taxonomist; black arrow), in other cases the imperfect classifier was wrong (red arrow), while in other cases still, the error-free classification is unknown due to lack of validation data. For example, the *Morphospecies 1* individual with no error-free classification must belong to a different column, but this species identity is unknown in the raw data. Ground beetle by Nikita Kozin, from the noun project

the few cases where the expert taxonomist could not identify a specimen to species level, we use their genus-level classification for the validation dataset.

Our dataset contains 5,865 individual specimens, 1,910 of which were identified by an expert taxonomist, and 62 species classified by the parataxonomist, 23 of which are morphospecies. Morphospecies identifications are unique to each NEON sampling location and year. We fit our model using all individuals and used no environmental covariates. Having both parataxonomist and expert taxonomist classifications complicates the use of NEON's carabid pitfall trap data (Figure 1). Only one study to date has been published using the NEON carabid pitfall trap data (Egli et al., 2020), but Egli et al. (2020) analyse only the subset of individuals that have expert taxonomist classifications. The classification-only model to which we compare our joint model can be fit only to the thinned dataset of verified samples, resulting in a loss of information.

2.2.2 | Model specification

We used informative priors for the species classification probability vectors $\theta_1, \ldots, \theta_K$ that placed higher probability density on the correct species classification. In the case of NEON beetle data, this is reasonable given the training that parataxonomists receive in beetle identification. Because all elements of each θ_k vector need to sum to one, and each element is bounded between 0 and 1, we used a Dirichlet prior: $\theta_k \sim \text{Dirichlet}(\alpha_k)$. We chose the Dirichlet concentration values α_k by comparing draws from the Dirichlet prior distribution to our prior intuition about imperfect classifier accuracy, making an assumption that there was a 65% chance that a species is correctly classified and some small probability that it is assigned to another specific class. Additionally, with smaller differences between the values of the Dirichlet prior, the model appeared not identifiable. This prior is informative particularly for components of Θ that have few observations (e.g. fewer than 80 observations on the diagonal, or fewer than 2 off-diagonal; see Table S2.1) such that the prior may be more informative than the data.

We used multivariate normal priors at the species and site level for initial occupancy, persistence, colonization and encounter rates. Correlated priors allow information sharing among parameters (Figure S3.1). The motivation for this stemmed from a prior expectation that these parameters could be related. For example, species with a higher encounter rate might be more likely to occur initially, persist or colonize new sites. Similar arguments could be made about relationships among site-level parameters. Each species is associated with a vector $\boldsymbol{\epsilon}_k$ of length 4, where $\epsilon_{k,1}$, $\epsilon_{k,2}$, $\epsilon_{k,3}$ and $\epsilon_{k,4}$ are species-specific adjustments on each of the four ecological parameters. The multivariate normal priors have a mean of zero and an unknown covariance matrix: $\boldsymbol{\epsilon}_k \sim \text{Normal}\left(\mathbf{0}, \sum^{(\alpha)}\right)$ with an Inverse-Wishart prior on $\sum$. Similarly, site-specific adjustments $\boldsymbol{\epsilon}_i$ were drawn from a different multivariate normal prior. These adjustments were added together on a transformed scale to compute initial occupancy, persistence, colonization and encounter rates, for example, $\text{logit}(\psi_{i,k,1}) = \epsilon_{i,1} + \epsilon_{k,1}$. A full model specification for the case study is available in Appendix S4 (Plummer et al., 2003).

To evaluate how the occupancy and encounter rate components of the full model informed classification probability estimates, we developed a reduced model that discards all information about occupancy and abundance, using just the true and imperfect species classifications to estimate the classification matrix Θ. This reduced model uses the same Θ prior as the joint model for consistency. Additionally, imperfect classifications are modelled as arising from a categorical distribution with a species-specific probability vector, as in the joint model. However, since occupancy and encounter rate information are discarded, true species classifications are not modelled and instead rely entirely on data. Thus, the reduced model is limited to the subset of data with verification. The comparison between full and reduced models reveals the extent to which occupancy and encounter rates inform classification probabilities. If there are no differences in the estimates of classification probabilities, then a two-stage model which first models misclassification and then passes the posterior of species classification on as a prior for an occupancy model should perform as well as the joint model in which the classification model is integrated with the occupancy model. However, we do not directly investigate the comparison in occupancy and encounter rate estimation between the joint and two-stage models.

All models were fit using JAGS with four chains of 20,000 iterations (Appendix S4, dclone, and R v4.0.2 (Plummer et al., 2003; R Core Team, 2020; Sólymos, 2010); output diagnostics were checked visually using traceplots and verifying that $\hat{R} < 1.1$ for estimates; and results were visualized with ggplot (Wickham, 2016).

2.3 | Simulation study

We conducted a simulation study to study the general behaviour of the model. We simulated 15,000 datasets with two species ($K = 2$) and three imperfect classifications ($\tilde{K} = 3$) while varying the fraction of verified samples. We expect model performance to decline as this fraction decreases.

Each dataset represents a single season of sampling with three surveys at each of 30 sites. The idea behind the three imperfect classifications is that two are the true species and one is a morphospecies (e.g. Species A, Species B and Morphospecies 1 in Figure 1). Every dataset was simulated using the priors used in the case study's joint classification-occupancy model (Appendices S3 and S4). The simulations use different Dirichlet concentration (α) values, and thus a different Θ prior, than the model used in the case study to accommodate a two-species dataset. We use a smaller dataset in the simulations compared to in the case study, specifically fewer species and a single season rather than multi-season. This dataset simplification was motivated for practical reasons of computational capacity, and we chose a similar number of species as was used in the simulations in Wright et al. (2020).

For a given simulated dataset, we know the true species classification for every sample. We fit both the joint

classification-occupancy and classification-only models to each dataset 100 times, where each time a different fraction of samples is verified. The 100 fractions for each dataset are randomly generated between 0.01 and 0.99. While this fraction range covers a wide variability in the verification effort, the results are sensitive to the choice of Θ prior, specifically the lower fractions would face estimation challenges (i.e. issues with parameter identifiability or MCMC convergence). If you arrange a dataset's 100 iterations in order from largest fraction of verified samples to smallest, when a sample's true species identity is removed in one iteration, it stays removed for subsequent iterations. We subset the samples this way to better isolate the effect that the validation fraction has on model results without the confounding effects of sample selection.

To evaluate model predictive performance, we first calculate validation metrics for the respective classification performance of the models by comparing estimated species identities to true species identities for samples with the species identity left out. However, we could not calculate validation metrics for the classification-only model since it cannot be fit without every sample having a verified species identity. For the purposes of calculating the validation metrics, we compare the true species identity, which is known in the simulations, to the distribution of median estimated species identities. That is, the estimated species identities are selected as the species that the median from the categorical posterior distribution (see Observation model: classification) lies on for each of the 1,200 chains for each iteration of every dataset. We evaluate how well the simulations predicted ecological parameters by measuring the coverage, difference between the estimated parameter value and true value, and width of 95% credible interval (CI) width for encounter rate (λ), occupancy (ψ) and classification probabilities (Θ_{full}) for the full model (joint classification-occupancy) and classification probabilities ($\Theta_{reduced}$) for the reduced model (classification-only).

3 | RESULTS

3.1 | Case study: Occupancy

Occupancy estimates varied across species and through time (Figure S2.2). No occupancy model was fit for the reduced, two-stage model, so there are not occupancy estimates to which to compare the full model's results. The joint occupancy model was designed to allow correlation between parameters across sites and species. Occupancy, growth and turnover rates also varied through time. Sites with high encounter rates tended to have low initial occupancy and colonization probabilities and high persistence probabilities (Figure S3.1). Furthermore, sites with high colonization rates tended to have high initial occupancy probabilities and low persistence probabilities. At the species level, we saw positive correlations among many of the model components, but in particular, species' encounter rate was positively correlated with species' initial occupancy, persistence and colonization rates (Figure S3.1). Species varied in their detection success by the imperfect classifier, from ones that were common and consistently identified correctly (e.g. *Calathus advena*) to ones that were not identified at all (e.g. *Dicheirotrichus mannerheimii*) but were caught by the expert taxonomist.

3.2 | Case study: Classification

The model yielded high probabilities of classification along the diagonal of the Θ confusion matrix where the true and imperfect identifications match (Figure 2). The parataxonomists, or imperfect classifiers in the NEON dataset, were trained in beetle taxonomy, so we built the model to favour the imperfect classifier by giving more weight in the Θ prior to diagonal values, making morphospecies classifications less probable. However, some species were just as or more likely to be identified as a morphospecies by the imperfect classifier than as the correct species. For example, the parataxonomist was more likely to classify *Pterostichus* (*Hypherpes*) sp. as morphospecies D13.2016.MorphBT than as the true species (see white arrows in Figure 2). However, no species had more than 3% probability (median) of being classified as another species (i.e. our model results indicate that the parataxonomist is most likely to identify a species either correctly or as a morphospecies). Samples with morphospecies classifications make up a sizeable portion of the carabid dataset, 812 out of the 5,865 total individuals identified by the imperfect classifier.

To evaluate the value added by informing the classification model with occupancy and encounter rates, we compared the full model to a reduced classification-only model that discards all information about occupancy and abundance. A difference between the models is their access to data. While the reduced model can use only a subset of the dataset, those samples with validation data, the advancement in the full model is that it allows the entire partially validated dataset to inform classification. Most θ_k probability vectors do not differ between the full and reduced model results. However, we see differences for a few species where there is less overlap in θ posteriors between the full and reduced models (e.g. Theta[*P.* (*Hypherpes*) sp., *P.* (*Hypherpes*) sp.] and Theta[*P. restrictus*, *P. restrictus*], Figure 3). These differences are found most notably for the abundant species. The full model yielded higher correct classification probabilities for the abundant species. Furthermore, the reduced model has larger 95% (CI) widths compared to the full model for many Θ indices (Figure 4). Thus, we find that a joint classification-occupancy model outperforms a two-stage model (classification, then occupancy) due to both the improvements in the modelling framework and the joint model's expanded access to the full dataset. Since the models being compared are fit to different sets of empirical data, we cannot quantify bias, which is important in model comparison.

We evaluated the performance of the full model's species classification by withholding some verified samples to see how well the model did in reproducing those true species identities. Validation metrics cannot be calculated for the reduced model since this model cannot be fit if true species identities are missing. We withheld a

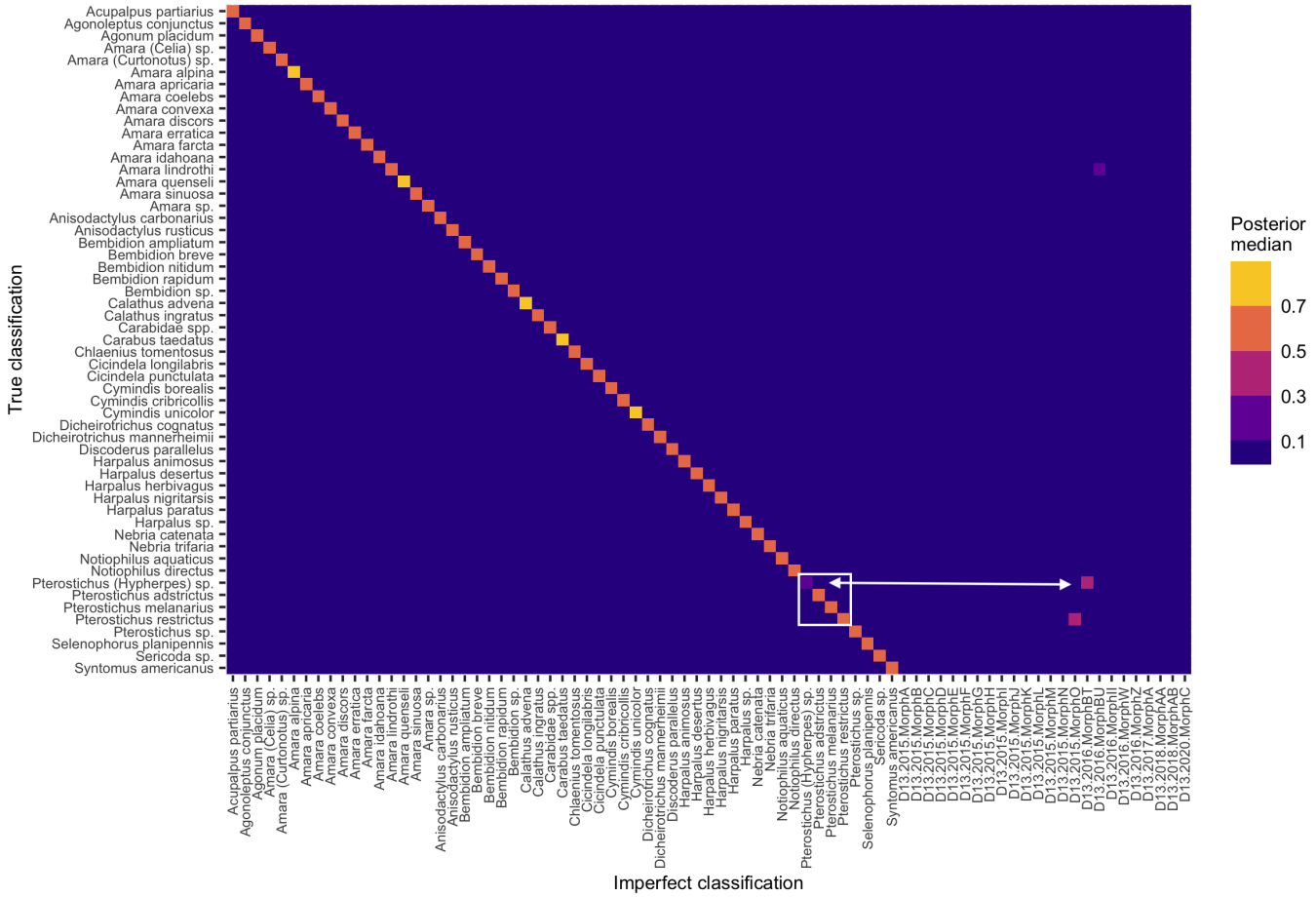

FIGURE 2 Θ confusion matrix for the full model (joint classification-occupancy model). Heat map of posterior median values where a value is interpreted as the probability that the species in that row is identified as a species in that column. Values along the diagonal are where the species is correctly identified. Values in each row sum to 1. The Θ posterior distributions for the 4 × 4 cells outlined in white are illustrated in Figure 3, along with the posteriors for the reduced (classification alone) and prior models. The white arrows point to an example where the imperfect classifier was more likely to classify a species as a morphospecies (D13.2016.MorphBT) than as the true species (*Pterostichus* (*Hypherpes*) sp.)

randomly selected 20% (382 individuals) of the true species identities (1,910 total). For the withheld samples, the classification accuracy was 89.9%, precision (for each imperfect classification, number correctly matched over total labelled as imperfect classification) 80.6%, and recall (for each species, number correctly matched over total samples of that species) 51.9%. Validation metric macro-averages are listed in Table 1.

3.3 | Simulation study

We found that model performance improved when fit to datasets with larger fractions of verified samples; however, the gains in model performance were modest. We illustrate model performance through model validation (Figure S5.1), coverage (Figure S5.3), difference between estimated parameter value and true value (Figure S5.2), and width of 95% credible intervals (Figure S5.4). In all of these metrics, we see that the gains in model performance are small or non-existent as the fraction of samples verified increases. We also found that the joint classification-occupancy model outperformed the classification-only model. Specifically, the joint model had higher precision (Figure 5). These results are specific to the dataset scenario specified for these simulations (e.g. 2 species, 1 morphospecies) and if parameter values were used for a dataset with higher species richness, the results would likely change.

4 | DISCUSSION

We developed a statistical approach that can be applied to datasets with imperfect observations that enhances multispecies classification by leveraging occupancy dynamics. This approach builds on recent work that integrates classification into occupancy models (Devarajan et al., 2020, and references therein) by evaluating the advantage of a joint classification-occupancy model, which allows imperfect classification categories to outnumber species; leverages individual-level confirmation data in a semi-supervised setting; and allows for the option to include covariates at the level of the

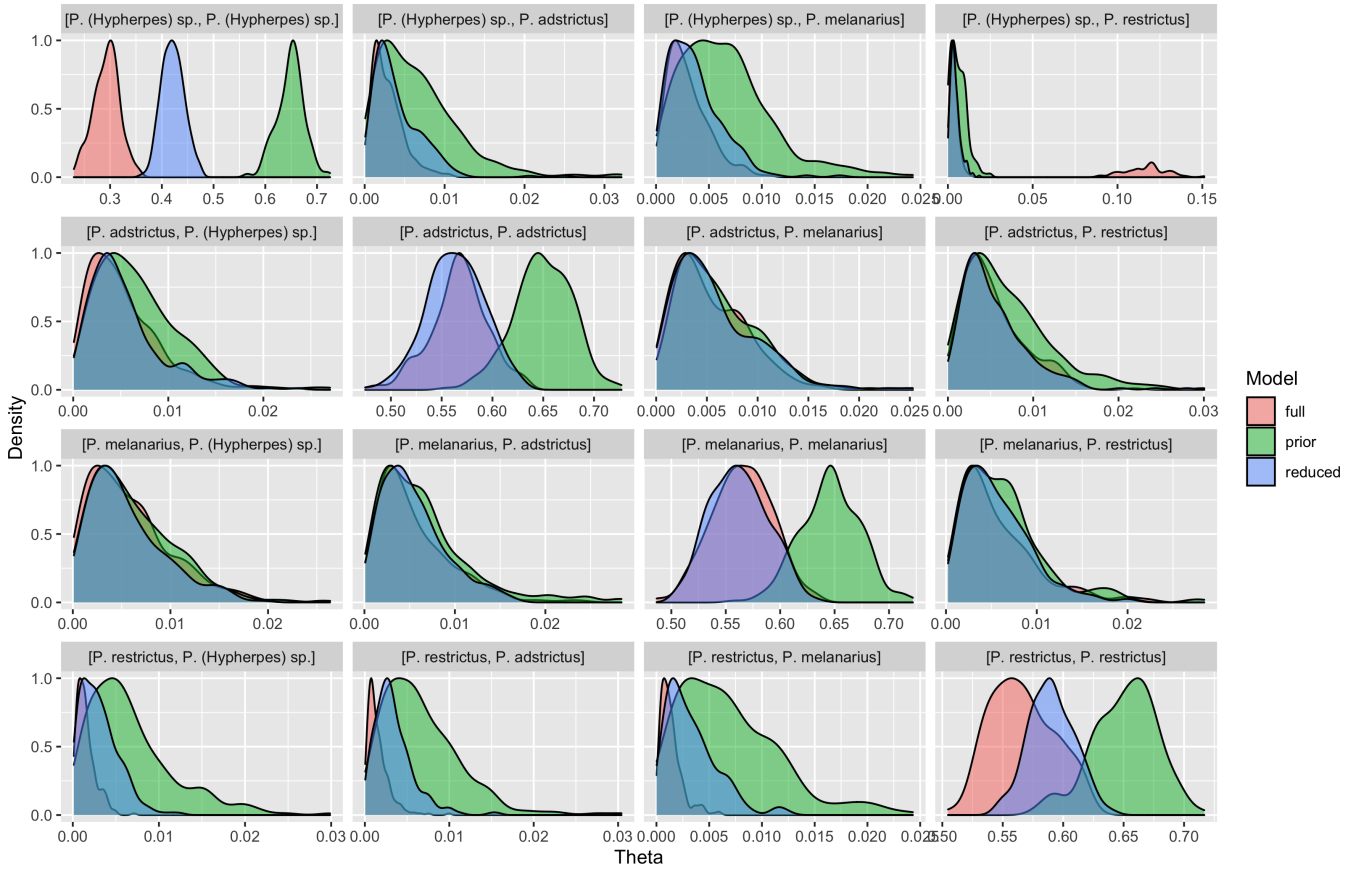

FIGURE 3 Comparison of θ density distribution between prior, full model posterior (classification with occupancy) and reduced model posterior (classification alone) for select θ confusion matrix indices for species in the *Pterostichus* genus. Each panel displays the probability distribution that the first species (true species identity) is imperfectly labelled as the second species (imperfect classification). For non-abundant species, θ_k probability distributions look like the second and third rows. In the second row, the three models' probability distributions overlap on the off-diagonal (where *P. adrictus* is imperfectly classified as other species) and overlap but differ from the prior along the diagonal (where *P. adrictus* is correctly classified as *P. adrictus*). In contrast, the top and bottom rows reflect probability distributions for abundant species. In the bottom row, the posterior distributions overlap and are narrower and smaller than the prior on off-diagonal values (where *P. restrictus* is imperfectly classified as other species). Along the diagonal, we see a difference in posterior probability distribution between the reduced and full models (where *P. restrictus* is correctly classified as *P. restrictus*), visualizing how the full and reduced models perform differently

individual/observation. Our probabilistic framework can be generalized to modelling abundance or other latent state variables that are estimated from multivariate count data with imperfect detection and classification (Chambert et al., 2016; Conn et al., 2013), as long as the data are classified at the individual level. For example, the $z * \lambda$ component of the encounter and observation models could alternatively be written as $p * N$ to model abundance. While analyses targeting species richness may be shielded to a certain extent from imperfect classification (Egli et al., 2020), any population- or community-level analysis with taxonomic specificity requires an understanding of classification uncertainty in the data. Our model provides a coherent statistical framework for ecological estimation in the presence of classification uncertainty.

False-positive and false-negative species classifications are inevitable in any field collection, caused by misclassification or non-detection and due to time and money constraints or imperfect classifiers (Hoekman et al., 2017; McClintock et al., 2010; Miller et al., 2012; Royle & Link, 2006). Misclassification may be caused by a number of extrinsic factors, including site- and survey-level covariates or observer error, the latter of which we focus on in the case study. Accounting for false identifications is important to reduce bias in occupancy dynamics estimated from multispecies biodiversity monitoring datasets (Chambert et al., 2015; McClintock et al., 2010; Miller et al., 2011; Miller et al., 2015) or in multi-state capture–recapture models (Pradel, 2005). Alternative models that account for false positives may consider data from only the focal species (Chambert et al., 2015) or from binary observations (Chambert et al., 2017). Like Wright et al. (2020), we use available counts from an imperfect classifier (Figure 1). However, we use all species detected, no matter how rare. By using a rectangular classification matrix that allows for propagation of taxonomic uncertainty for multispecies datasets where imperfect classifications outnumber species (e.g. unknown, morphospecies, to the family level; Figure 2), we remove a limitation that previous occupancy modelling methods have used (Chambert et al., 2018; Wright et al., 2020). For example, although the model priors favour imperfect classifier accuracy, the

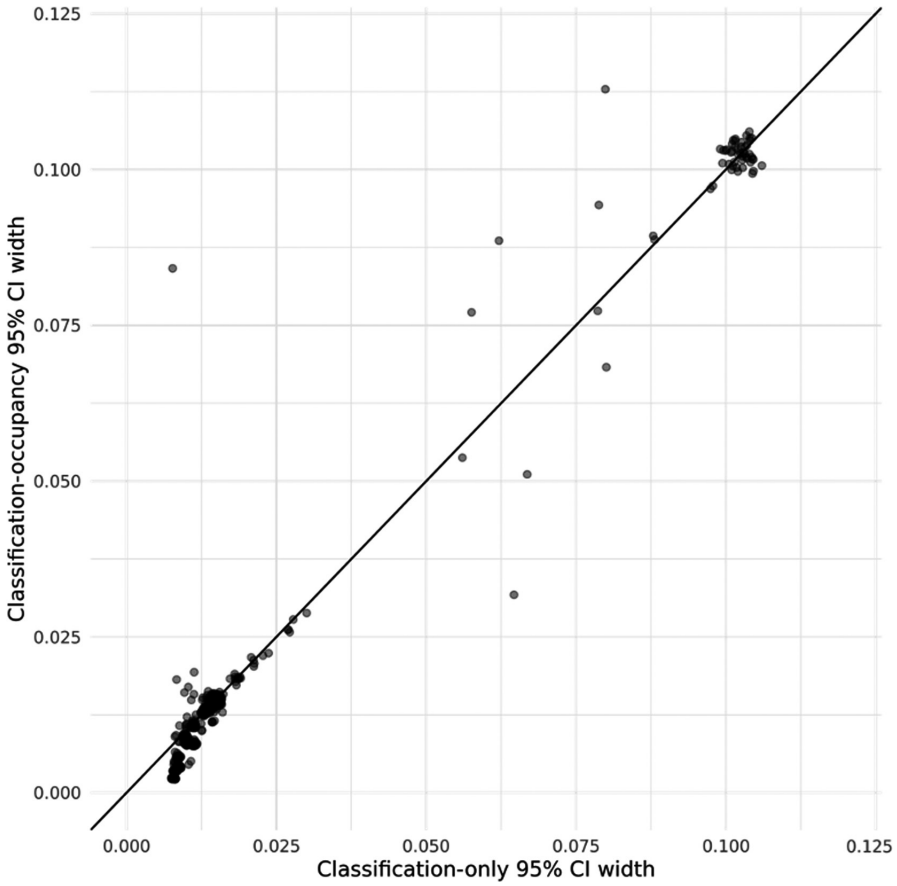

FIGURE 4 Comparison of precision between θ posterior 95% credible intervals (CI) of joint classification-occupancy (full) model and classification-only (reduced) model. The full model is more precise for points below the line, indicating that the reduced model has a larger CI than the full model for that species. The clustering pattern is driven by whether the θ value is close to zero or not. Mostly along the diagonal, larger θ values predict that the imperfect and verified classifications match, which yields larger CI widths, whereas most off-diagonal θ values are close to 0 and have strong confidence, or smaller 95% CI width (Figures 2 and 3)

model found likely species matches for a couple of morphospecies that were abundant in the data (e.g. D13.2015.MorphO, D13.2016.MorphBT; Figure 2).

Our model is semi-supervised and makes use of data at the individual level. Whereas alternative models use data pooled at the site or visit level (Chambert et al., 2018; Wright et al., 2020), our model leverages the rich, individual-level information to reveal which species are commonly mistaken by the imperfect classifier and how often, allowing species counts to inform classification (Chambert et al., 2017). Verified individuals can be used as partially observed occupancy data in our semi-supervised model. Our model could be expanded to include observation-level covariates (e.g. sample confidence), which is an extension that our individual-level, occupancy-detection model can achieve that a count-detection model cannot. The case study in Wright et al. (2020) classifies bat calls to species at the site and visit level. All calls are classified by an automated, imperfect classifier and only the subset of calls that are classified manually by an error-free classifier are used to fit their model, though the authors provide a model extension if detections are unable to be confirmed. The Wright et al.'s (2020) model is analogous to our joint model if the data were aggregated to the site and visit level.

Ours is the first model to consider how occupancy and encounter rates contribute to improving species classification (although Augustine et al. (2020) developed a capture–recapture model for improving individual classifications). We found that our joint classification-occupancy model outperformed a classification-only model that disregarded occupancy dynamics and could use only a supervised subset of the data in estimating imperfect classification. Specifically, the joint model yielded more precise estimates (Figures 4 and 5). Accuracy of classification estimates was not assessed for either model because true misclassification probabilities are not known for the case study data. While there was large agreement between the confusion matrix Θ of both models, the full model had higher probability estimates for abundant species (Figure S2.1).

TABLE 1 Validation metric macro-averages for joint classification-occupancy model. Accuracy is the number of species classifications correct over the total number of classifications; precision is the macro-average of for each imperfect classification category, the number correctly matched over the total labelled as that category; recall is the macro-average of for each species, the number correctly matched over total samples of that species; the F1 score combines precision and recall into a single number; and the holdout log-likelihood is the log-likelihood for the holdout data subset. See associated code for formulas used

Metric	Value
Accuracy	0.899
Precision	0.806
Recall	0.519
F1 score	0.852
Holdout log-likelihood	−249

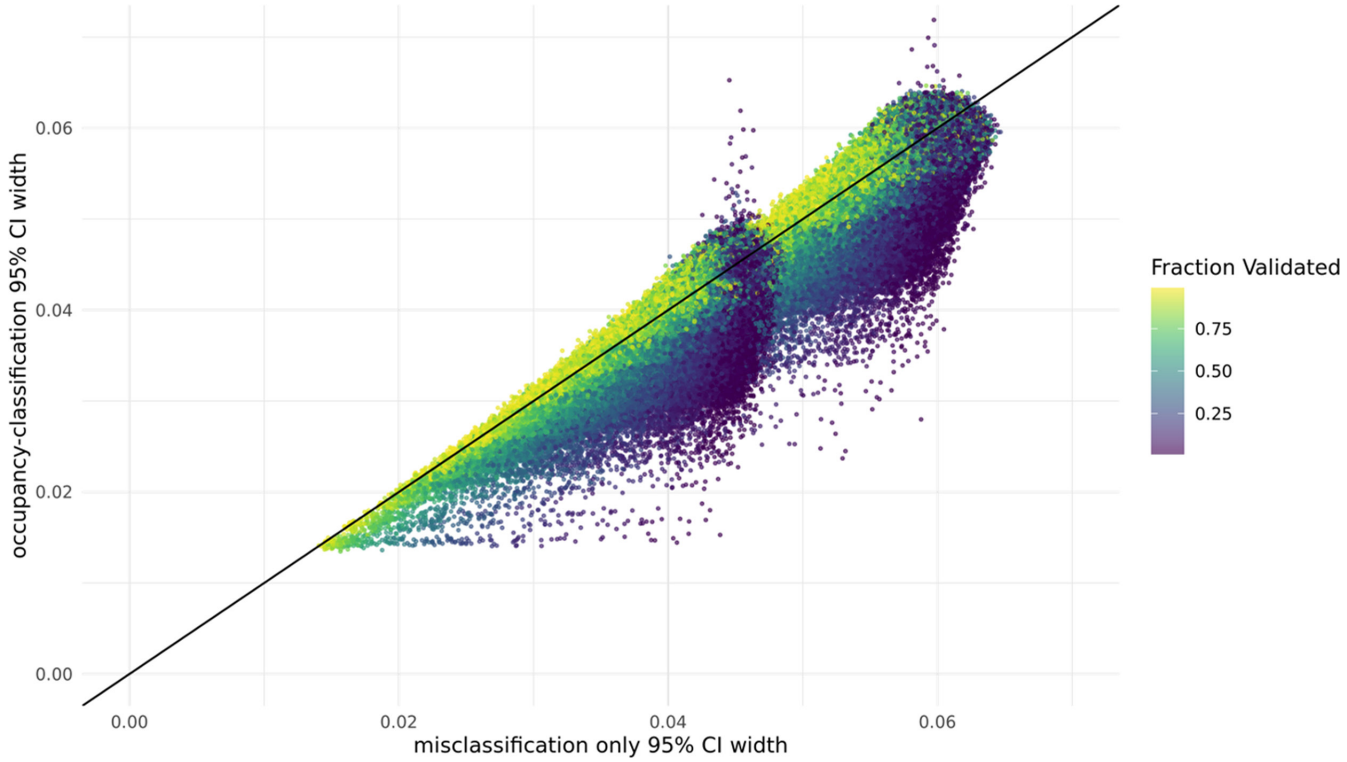

FIGURE 5 Comparison of precision between θ posterior 95% credible intervals (CI) of joint classification-occupancy (full) model and classification-only (reduced) model for simulation results. The full model is more precise for points below the line, indicating that the reduced model has a larger CI than the full model for that θ index in a given dataset. The clustering pattern is driven by whether the θ value is close to zero or not. Mostly along the diagonal, larger θ values predict that the imperfect and verified classifications match, which yields larger CI widths, whereas most off-diagonal θ values are close to 0 and have strong confidence, or smaller 95% CI width

The simulations explore how model behaviour changes with the fraction of samples verified. Error-free classification of samples is often costly and time-intensive, which makes understanding the added value of more verification useful for any researchers conducting species surveys on a budget. From our simulations, we found that a higher fraction of verified samples yielded better model performance, though the difference was modest (Appendix S5). For example, the difference between the classification accuracy for a dataset with 90% of its samples verified versus a dataset with almost none of its samples verified is roughly 5% (Figure S5.1). While this difference is small, it can likely be explained by the informative Dirichlet prior paired with a dataset with few species. While the validation metrics (Figure S5.1) illustrate some improvement with more samples validated, this trend is mostly unnoticeable for ecological parameter estimates from the simulations (Figures S5.3, S5.2 and S5.4). This modest difference in model performance also highlights the balance that is to be struck between budgeting for more sample verification to meet desired research goals and getting fewer samples verified but still yielding reasonable, though not optimal, model performance. The simulations presented are for a two-species case, but model performance may look different for scenarios with more species. Researchers that have an idea of how many species they will encounter in a survey and how many samples they can budget for verification can adapt these simulations to inform how many samples they choose to be verified.

Our model has a number of limitations. For one, the joint classification-occupancy model takes more computing power to fit than the classification-only model. For the NEON data used in the case study, the imperfect classifier had high agreement with verified classifications, and we reflected this in the model by using informed Dirichlet priors that offer the imperfect classifier high accuracy. However, the model results may vary when fit to datasets with a less reliable imperfect classifier. The extent to which imperfect classifier accuracy affects model behaviour could be addressed through simulations. A model assumption is that samples were selected at random for verification. However, the dataset used in the case study gave verification priority to samples that the imperfect classifier could not identify to species level (National Ecological Observatory Network, 2021). Violating a model assumption by having samples that are low in classification confidence preferentially selected for verification introduces bias into the confusion matrix parameter estimates, likely causing misclassification probabilities to be overestimated. In other cases, like when data collection methodology changes in long-term surveys, simulations could inform which types of samples should be prioritized for verification to yield desired model results.

We tried various iterations of the model before coming to the final semi-supervised, individual-level model. While an aggregated data approach theoretically takes less processing time, we found the model fit to aggregated data would either not converge or struggled

with identifiability, yielding multimodal posteriors for θ. Changing the Dirichlet priors to favour imperfect classifier accuracy helped but did not eliminate the problem. In the case study, we investigate differences in classification performance between the full and reduced models and do not directly investigate and compare performance of the model in estimating ecological parameter estimates. Future work could more explicitly address false positives by informing θ priors with species commonly misidentified by the imperfect classifier or by inducing sparsity in the θ matrix through setting certain priors to 0.

Large-scale, long-term biodiversity surveys are critical to inform land management and conservation policy (Hughes et al., 2017) and require accurate species classifications to achieve conservation objectives. This probabilistic approach can model species occupancy while accounting for imperfect detection and classification. Considering misclassification is even more important when making inferences about temporal transition (i.e. extinction, colonization) than for occupancy because misclassification in either of the focal time periods (t or $t + 1$) can produce bias. Innovations in occupancy models, in general, are rapidly being made to consider an expanding variety of study systems and experimental designs (Bailey et al., 2014) and most of these approaches rely on observation-level, ground-truthed verification data for model training and validation. The proposed model can be extended to enhance occupancy inferences made from citizen science surveys (Sauer et al., 2017); long-term, multi-PI surveys where methodologies may vary through time or across sites; automated imperfect classifiers (i.e. machine learning algorithm) applied to large volumes of open data (e.g. satellite or airborne remote sensing, camera traps); or research scenarios where data provenance is limited, which makes propagating classification uncertainty important. Future work could explore through simulation ways that model parameters affect behaviour (e.g. effects of different weights on the informed Dirichlet prior, imperfect classifier accuracy or sample prioritization for verification) and use observation-level covariates to enhance estimation of classification probabilities and ecological parameters.

ACKNOWLEDGEMENTS

The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle. This material is based in part upon work supported by the National Science Foundation through the NEON Program. We thank G. Vagle for taking part in the conception of this project and J. Coulombe for her graphical design assistance. The work was supported by the CU Boulder Grand Challenge investment in Earth Lab. AIS was supported as a GRA at Earth Lab for work on this project. Any use of trade, product or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

AUTHORS' CONTRIBUTIONS

A.I.S., C.L.T. and M.B.J. conceived the project idea; A.I.S., J.A.R. and M.B.J. designed the methodology; A.I.S. curated the data; A.I.S. and M.B.J. analysed the data; A.I.S. and M.B.J. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13858.

DATA AVAILABILITY STATEMENT

Carabid data are publicly accessible through the NEON data portal at https://data.neonscience.org/data-products/DP1.10022.001 (National Ecological Observatory Network, 2021). Code for data cleaning and analysis is archived on Zenodo at https://doi.org/10.5281/zenodo.6395234 (Spiers et al., 2022) and also available at https://github.com/annaspiers/NEON-NIWO-misclass.

ORCID

Anna I. Spiers https://orcid.org/0000-0003-3517-1072
J. Andrew Royle https://orcid.org/0000-0003-3135-2167
Christa L. Torrens https://orcid.org/0000-0002-1036-7645
Maxwell B. Joseph https://orcid.org/0000-0002-7745-9990

REFERENCES

Augustine, B. C., Royle, J. A., Linden, D. W., & Fuller, A. K. (2020). Spatial proximity moderates genotype uncertainty in genetic tagging studies. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(30), 17903–17912.

Bailey, L. L., MacKenzie, D. I., & Nichols, J. D. (2014). Advances and applications of occupancy models. *Methods in Ecology and Evolution*, *5*(12), 1269–1279. https://doi.org/10.1111/2041-210X.12100

Chambert, T., Grant, E. H. C., Miller, D. A., Nichols, J. D., Mulder, K. P., & Brand, A. B. (2018). Two-species occupancy modelling accounting for species misidentification and non-detection. *Methods in Ecology and Evolution*, *9*(6), 1468–1477. https://doi.org/10.1111/2041-210X.12985

Chambert, T., Hossack, B. R., Fishback, L., & Davenport, J. M. (2016). Estimating abundance in the presence of species uncertainty. *Methods in Ecology and Evolution*, *7*(9), 1041–1049.

Chambert, T., Miller, D. A., & Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, *96*(2), 332–339. https://doi.org/10.1890/14-1507.1

Chambert, T., Waddle, J. H., Miller, D. A., Walls, S. C., & Nichols, J. D. (2017). A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing. *Methods in Ecology and Evolution*, *9*(3), 560–570. https://doi.org/10.1111/2041-210X.12910

Conn, P. B., McClintock, B. T., Cameron, M. F., Johnson, D. S., Moreland, E. E., & Boveng, P. L. (2013). Accommodating species identification errors in transect surveys. *Ecology*, *94*(11), 2607–2618.

Devarajan, K., Morelli, T. L., & Tenan, S. (2020). Multi-species occupancy models: Review, roadmap, and recommendations. *Ecography*, *43*, 1612–1624. https://doi.org/10.1111/ecog.04957

Dorazio, R. M., & Royle, J. A. (2005). Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, *100*(470), 389–398. https://doi.org/10.1198/016214505000000015

Egli, L., LeVan, K. E., & Work, T. T. (2020). Taxonomic error rates affect interpretations of a national-scale ground beetle monitoring program at national ecological observatory network. *Ecosphere*, *11*(4), e03035. https://doi.org/10.1002/ecs2.3035

Hoekman, D., LeVan, K. E., Ball, G. E., Browne, R. A., Davidson, R. L., Erwin, T. L., Knisley, C. B., LaBonte, J. R., Lundgren, J., Maddison, D. R., et al. (2017). Design for ground beetle abundance and diversity sampling within the national ecological observatory network. *Ecosphere*, *8*(4), e01744. https://doi.org/10.1002/ecs2.1744

Hughes, B. B., Beas-Luna, R., Barner, A. K., Brewitt, K., Brumbaugh, D. R., Cerny-Chipman, E. B., Close, S. L., Coblentz, K. E., De Nesnera, K. L., Drobnitch, S. T., et al. (2017). Long-term studies contribute disproportionately to ecology and policy. *Bioscience*, *67*(3), 271–281. https://doi.org/10.1002/ecs2.1744

Iknayan, K. J., Tingley, M. W., Furnas, B. J., & Beissinger, S. R. (2014). Detecting diversity: Emerging methods to estimate species diversity. *Trends in Ecology & Evolution*, *29*(2), 97–106. https://doi.org/10.1016/j.tree.2013.10.012

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*(8), 2248–2255.

McClintock, B. T., Bailey, L. L., Pollock, K. H., & Simons, T. R. (2010). Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology*, *91*(8), 2446–2454. doi:10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

Miller, D. A., Bailey, L. L., Grant, E. H. C., McClintock, B. T., Weir, L. A., & Simons, T. R. (2015). Performance of species occurrence estimators when basic assumptions are not met: A test using field data where true occupancy status is known. *Methods in Ecology and Evolution*, *6*(5), 557–565. https://doi.org/10.1111/2041-210X.12342

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, *92*(7), 1422–1428. https://doi.org/10.1890/10-1396.1

Miller, D. A., Weir, L. A., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Simons, T. R. (2012). Experimental investigation of false positive errors in auditory species occurrence surveys. *Ecological Applications*, *22*(5), 1665–1674. https://doi.org/10.1890/11-2129.1

National Ecological Observatory Network. (2021). *Data products: Neon. dp1.10022.001*. Retrieved from https://data.neonscience.org/data-products/DP1.10022.001

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, *20*(5), 561–576. https://doi.org/10.1111/ele.12757

Plummer, M. (2003). Jags: A program for analysis of Bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (vol. *124*, pp. 1–10).

Pradel, R. (2005). Multievent: An extension of multistate capture–recapture models to uncertain states. *Biometrics*, *61*(2), 442–447.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, *87*(4), 835–841. https://doi.org/10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2

Sauer, J. R., Niven, D. K., Hines, J. E., Ziolkowski, D. J., Pardieck, K. L., Fallon, J. E., & Link, W. A. (2017). *The north American breeding bird survey, results and analysis* (pp. 1966–2015). Retrieved from https://www.mbr-pwrc.usgs.gov/bbs/bbs.html

Sólymos, P. (2010). Dclone: Data cloning in R. *The R Journal*, *2*(2), 29–37. https://doi.org/10.32614/RJ-2010-011

Spiers, A., Torrens, C., Vagle, G., & Joseph, M. (2022). *annaspiers/NEON-NIWO-misclass: v2.1.1*. https://doi.org/10.5281/zenodo.6395234

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., Teton, B., Beasley, J. C., Schlichting, P. E., Boughton, R. K., Wight, B., Newkirk, E. S., Ivan, J. S., Odell, E. A., Brook, R. K., … Miller, R. S. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, *10*(4), 585–590. https://doi.org/10.1111/2041-210X.13120

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. Retrieved from https://ggplot2.tidyverse.org

Wright, W. J., Irvine, K. M., Almberg, E. S., & Litt, A. R. (2020). Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. *Methods in Ecology and Evolution*, *11*(1), 71–81. https://doi.org/10.1111/2041-210X.13315

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Spiers, A. I., Royle, J. A., Torrens, C. L. & Joseph, M. B. (2022). Estimating species misclassification with occupancy dynamics and encounter rates: A semi-supervised, individual-level approach. *Methods in Ecology and Evolution*, 13, 1528–1539. https://doi.org/10.1111/2041-210X.13858